Are Signature Proteins the Key to Evaluating Eukaryotic Phylogeny?

Jian Han[1]*, Lesley J. Collins[2], Patrick J. Biggs[1], W. Timothy White[1], David Penny[1]

[1]Institute of Molecular BioSciences, Massey University, Palmerston North, New Zealand

[2]Universal College of Learning, Palmerston North, New Zealand

*Corresponding author: Jian Han, jian272@hotmail.com

Abstract

Eukaryotic Signature Proteins (ESPs) are proteins that delineate the eukaryotes from the archaea and bacteria. They have no readily recognisable homologues in any prokaryotic genome, but their homologues are present in all main branches of eukaryotes. Since ESPs are conserved in all eukaryotes, they are considered to be ancient proteins with a slow rate of evolution. These properties make them theoretically good candidates to analyse the phylogenetic relationships of eukaryotic species.

This study examines the possibilities of using ESPs as a special group of proteins in deep phylogenetic analysis. After concatenation of ESP sequences, a phylogenetic tree of 15 mammalian species, and another phylogenetic tree including many divergent eukaryotes was generated to examine how ESPs perform. The latter tree is the longest concatenated sequences to built tree for eukaryotes to date. ESP performed very well for the mammal phylogenetics. The deep eukaryotic tree included some deeply divergent branches but has also showed promising results.

Keywords

Eukaryotic Signature Protein (ESP); Phylogenetics; Giardia Lamblia; Eukaryote; Mammal

Introduction

In recent years, technological advances have seen the phylogeny of deep eukaryotes change frequently, primarily through the application of phylogenetics and genomic sequencing. Phylogenetic relationships between distant eukaryotic species were initially performed based on single ubiquitous genes such as 18s rRNA, elongation factors and tubulins (Hashimoto et al. 1994). Using 18s rRNA has the advantage of being easy to amplify with PCR due to highly conserved flanking regions allowing for the use of universal primers (Meyer et al. 2010). However, the downside of using 18s rRNA is that the accuracy can also be lowered by factors such as mutational saturation, unequal mutation rates and rapid evolutionary radiation (Philippe and Adoutte 1998). It cannot resolve nodes at all taxonomic levels and its efficacy varies considerably among clades (Abouheif, Zardoya, and Meyer 1998), so we can end up with a lack of resolution (stochastic error) because of a low number of informative sites, and systematic error in tree estimation caused by model violations. There are also problems related to long branch attraction (LBA, a phenomenon when highly divergent lineages are grouped together, regardless of their true evolutionary relationships) (Philippe 2000).

Due to an increasing number of genomes available, it is now possible to compute gene trees for many different genes. Researchers can now attempt to obtain a more reliable species tree by phylogenomic approach, concatenating sequences of multiple genes and then building a species tree from the concatenated sequences (Burki et al. 2007; Hampl et al. 2009). The most difficult trees to reconstruct are those that bring together diverse species. The eukaryotic cell is suggested to have arisen 1850 million years ago (Knoll et al. 2006). The current system classifies eukaryotes into five supergroups based on molecular and morphological/cell-biological evidence (Keeling et al. 2005; Keeling 2007). These are Unikonta (note: some literature subdivides this supergroup into Opisthokonta and Amoebozoa (Simpson 2003)), Plantae (aka Archaeplastida), Rhizaria, Chromalveolata, and Excavata. However, there is still debate surrounding this classification system, because the earliest eukaryotic divergences (i.e. the root of the tree) are unresolved at present (Keeling 2007). In addition, molecular phylogeny has not yet provided clear evidence that Excavata and Chromalveolata are monophyletic (i.e. all members are derived from a unique common ancestor) (Parfrey et al. 2006; Hampl et al. 2009).

Hampl *et al.* (Hampl et al. 2009) studied the monophyly of the Excavata supergroup by concatenating 143 gene sequences from 48 species

from all five supergroups including 19 excavates and concluded that Excavata forms a monophyletic supra-kingdom-level group. However, even with the removal of long-branch gene sequences (removal of individual genes which have accumulated large numbers of changes), they could only obtain a somewhat unconvincing bootstrap value of 54%. With long-branch taxa removing the support increased to 90%, but *Giardia* and *Trichomonas* were two of the taxa that they removed, and there were no diplomonads and parabasalids (two major branches of Excavata (Simpson, Inagaki, and Roger 2006) representatives in their tree. The genes chosen in Hampl *et al.*'s studies were a random selection of available genes as long as the gene sequence was available in the species of interest (Hampl et al. 2009). Eukaryotic Signature Proteins (ESPs), a special set of proteins, are conserved throughout eukaryotes (Hartman and Fedorov 2002; Kurland, Collins, and Penny 2006; Han and Collins 2012), and thus theoretically they could outperform other random selections of proteins to determine the phylogenetic relationship of eukaryotes. Hence this study is carried out to test this hypothesis.

ESPs, proteins that delineate the eukaryotes from the archaea and bacteria, have no recognisable homologues in any prokaryotic genome, but their homologues are present in all main branches of eukaryotes (Hartman and Fedorov 2002) and thus being theorised to be ancient proteins with a slow and consistent evolving rate. The fact that ESPs are conserved in all groups of eukaryotes minimises the amount of "missing data" from any organism when dealing with eukaryotic phylogenies. Some species, such as *Giardia lamblia* (from here on *Giardia*), have a large number of genes which appear more similar in sequence to bacterial genes, thus biasing the position of *Giardia* in eukaryotic phylogenetic studies (Nixon et al. 2002; Andersson et al. 2003; Morrison et al. 2007). ESPs can prevent this scenario because the proteins with bacterial homologues are removed. These properties have made them in theory at least, good candidates to analyse the phylogenetic relationships of eukaryotic species. A previous set of ESPs for *Giardia* was calculated before the sequencing of many eukaryotic genomes (Hartman and Fedorov 2002), but for this study a recalculated ESP dataset was used (Han and Collins 2012).

This study examines the possibilities of using ESPs as a special group of proteins in phylogenetic analysis. The phylogenetic relationship of 15 mammalian species was first analysed before deep phylogenetic analysis. With well known fossil and morphological evidence of mammalian phylogeny that our results can be compared, we can thus assess how accurately ESPs perform in less deeper phylogenetic analysis. After that the phylogenetic relationship of 18 eukaryotic species, including some divergent species such as *Giardia*, *Dictyostelium* and *Phytophthora*, were analysed using ESPs. The aim of this research is not focused on the actual phylogenetic relationship between the supergroups of eukaryotes, but instead to investigate how good ESPs are as candidates for further phylogenetic research.

Methods

ESPs were calculated as per Han and Collins 2012 (Han and Collins 2012). This can be summarised as follows: the chosen starting organism of *Giardia lamblia* with 4889 annotated proteins. *Giardia* proteins that had homologues in any of the 28 bacterial and 12 archaeal species were then discarded; then proteins that did not have homologues in any of the 17 eukaryotic species (Table 1) were removed. BLAST hits with a bit-score ≥ 55 were considered as "homologues". Finally 267 *Giardia lamblia* eukaryotic signature proteins (ESPs) were obtained (Han and Collins 2012).

TABLE 1 EUKARYOTIC SPECIES USED FOR THE CALCULATION OF ESPS AND PHYLOGENETIC ANALYSIS OF EUKARYOTES

Species	Common names	Supergroup
From Ensembl		
Aedes aegypti	Yellow fever mosquito	Opisthokonta
Caenorhabditis elegans	-	Opisthokonta
Canis familiaris	Dog	Opisthokonta
Ciona intestinalis	Sea squirt	Opisthokonta
Danio rerio	Zebrafish	Opisthokonta
Drosophila melanogaster	Fruitfly	Opisthokonta
Gallus gallus	Chicken	Opisthokonta
Mus musculus	Mouse	Opisthokonta
Tetraodon nigroviridis	Pufferfish	Opisthokonta
Xenopus tropicalis	Western clawed frog	Opisthokonta
From Dictybase		
Dictyostelium discoideum		Amoebozoa
From Broad Institute		
Neurospora crassa	-	Opisthokonta
Phytophthora infestans	-	Chromalveolata
From NCBI		
Arabidopsis thaliana	-	Plantae
Oryza sativa	Rice	Plantae
Schizosaccharomyces pombe	-	Opisthokonta
From Aspgd		
Aspergillus nidulans	-	Opisthokonta
From Giardiadb		
Giardia lamblia	Giardia	Excavata

For each *Giardia* ESP, the highest scored homologue from each of 17 eukaryotic organisms (Table 1) was recovered using a custom Perl script. The original sequences from *Giardia* as well as its homologues from the 17 eukaryotic organisms were then aligned using ClustalX version 2.0.11 (Larkin et al. 2007). This procedure was then repeated for each of the 267 *Giardia* ESPs. All alignments were imported into software Geneious Pro version 5.0.4 (Drummond AJ et al. 2011) for further phylogenetic analyses. Geneious plug-in PHYML (Guindon and Gascuel 2003) was used to create maximum likelihood (ML) trees for each alignment with ten bootstraps performed for each tree. This number (ten) was relatively low but a compromise for the time taken to build 267 trees. These trees were built only for primary analyses and a high number of bootstraps was not essential at this stage.

The 267 trees were then manually examined to determine the best individuals for phylogenetic studies. By using *a priori* phylogenetic knowledge, the trees were divided into three groups: Group A (excellent), Group B (acceptable) and Group C (unsuitable) based on the more detailed definitions below:

Group A trees have all animals (*Aedes, Caenorhabditis, Canis, Ciona, Danio, Drosophila, Gallus, Homo sapiens, Tetraodon, Xenopus*) in one clade, all plants (*Arabdopsis* and *Oryza*) in one clade and all fungi (*Aspergillus, Neurospora* and *Schizosaccharomyces*) in one clade, the bootstrap value of any of these three clades has to be no less than 70 percent. Placement of other organisms *Phytophthora*, *Dictyostelium* and *Giardia* is of less priority so long as they do not appear inside the three clades mentioned above. This is because the phylogenetic ordering of these three longer-branching organisms is less clear. Group A trees were considered of excellent quality, because the animals, fungi and plants are expected to be monophyletic.

Group B trees have only one or two species misplaced within the major clades, or if there are low bootstrapping values for the three clades mentioned above (even if the topology of the tree fulfils all the requirement of Group A).

Group C trees contain so called "star" trees, in which all major branches originate from a single point, implying that all of these branches are unresolved. This group also contains trees having more than two clearly misplaced species (e.g. animal species grouped with fungi). Any tree that displayed properties prevents them from falling into the Group A and Group B was placed in this group.

After this filtering procedure, Group C (unsuitable) trees were discarded, and the subsequent analyses were performed using the Group A (excellent) and Group B (acceptable) trees only (a total of 140 trees from the original 267).

For the mammal phylogenetic analysis, complementary DNA sequences were used instead of protein sequences. Fifteen mammalian species from Ensembl, all with at least 6X genome coverage were analysed (Table 2). The ***Anolis carolinensis*** (Anole Lizard) genome was also downloaded to serve as an outgroup for this study. Each of the 50 Group A and 90 Group B ESPs were BLASTed against each mammal, and the hit with highest bit-score was recorded from each organism. The sequences were aligned and then the alignments were concatenated. After a model test, the ML tree was built using Hasengawa-Kashino-Yano 1985 (HKY85) substitution matrix (Hasegawa, Kishino, and Yano 1985). This tree was bootstrapped 200 times.

For the deep eukaryotic tree, the phylogenetic relationship among the 18 organisms used for the initial alignments were analysed. Group A and Group B protein sequences belonging to the same organism were concatenated, and a new ML tree was built on these linked ESP sequences. After a model test was performed, the WAG substitution matrix, with a four–discrete-category gamma approximation (''WAG + Γ4 model'') was chosen to be amino acid substitution matrix. The resulting tree was bootstrapped 100 times.

Results

ML Trees of ESP

The phylogenetic trees were expected to have all the animals forming a clade of their own, as do the three fungi and plant species. The divergent species (*Giardia*, *Dictyostelium* and *Phytophthora*) formed long branches. However, trees with unexpected shapes can be formed if a wrong paralogue (i.e. after a gene duplication, one copy of the gene may change function as it accumulates mutations) was used for tree construction. If the correct paralogue (i.e. the original copy of the gene that has retained the original function) is used then the tree should display the true phylogenetic relationship. One of the ESPs showing this misplaced paralogue effect was the 26S proteasome non-ATPase regulatory subunit 7 (GL50803_7896) (Fig. 1).

Clearly in this situation, *Gallus* has been misplaced

into the same clade with *Giardia*, with 99% bootstrap value (Fig. 1A). This protein has many paralogues in *Gallus* with the best match being ENSGALP00000008530 (Protein A) with a bit-score of 65.5. The other strong match was ENSGALP00000000999 (Protein B) with a slightly less bit-score of 65.1. In the default tree generating procedure, Protein A was used as the *Gallus* protein because of its higher bit-score. When Protein B was used as the *Gallus* protein, a different tree was generated, and *Gallus* was placed back to the animal clade where it clearly belongs (see Fig. 1B). For some trees, the obvious misplacing mistakes could not be fixed by using a different paralogue, e.g. GL50803_15339 (Adaptor protein complex large chain subunit BetaA), no matter which paralogue was used, *Ciona* remains as a long branch (Fig. 2). This may indicate that all the paralogues may have evolved rapidly.

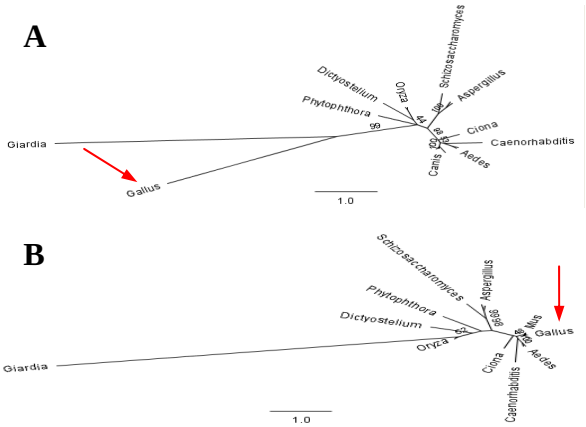

FIG. 1 UNROOTED ML TREE OF ORTHOLOGUES OF GL50803_7896 (26S PROTEASOME NON-ATPASE REGULATORY SUBUNIT 7) FROM DIFFERENT SPECIES SHOWING EFFECT OF INCLUDING AN INCORRECT GENE GALLUS PARALOGUE. BOOTSTRAPPING VALUES ARE SHOWN ON BRANCHES, AND THE SCALE BAR IS NUMBER OF SUBSTITUTIONS PER SITE.

A.ENSGALP00000008530 (Protein A) was used as Gallus orthologue. Gallus (indicated by the arrow) has been grouped with Giardia when grouped with other metazoans.

B.ENSGALP00000000999 (Protein B) was used as Gallus orthologue. Using the correct homologue, Protein B has placed Gallus back in the animal clade and displayed a greater distance between the Giardia protein and its homologues (Gallus branch is indicated by the arrow).

To find trees that might be susceptible to the above scenario, i.e. having the incorrect paralogues from species causing that incorrect phylogenetic relationship to be portrayed, the 267 trees each manually checked and were divided into three groups (see method section for detailed explanation of the three groups) based on the topology and bootstrap support of the tree:

- Group A contained what considered to be 50 excellent trees, each of which has all animals, fungi and plants into three separate clades, showing that these three are clear monophyletic with bootstrapping value no less than 70 percent.
- Group B contained 90 trees each having only one or two species being misplaced, or with low bootstrapping values for the three clades mentioned above even the topology of the tree is good.
- Group C contained 127 trees which are considered trees not very useful for phylogenetic study, and misplaced a large number of species. Group C trees were not used for the subsequent analyses.

Ribosomal biogenesis ESPs and ESPs found in vacuole, ER and Golgi have a tendency to fall into Group A and B. On the contrary, a high proportion of cytoskeletal ESPs fell into Group C. This might be due to the fact that actins and tubulins have many paralogues. ESPs of the signalling pathways (notibally kinases and phosphotases) were also most likely to be found in Group C containing ESPs that are short sequences and there is not enough sites to produce meaningful results.

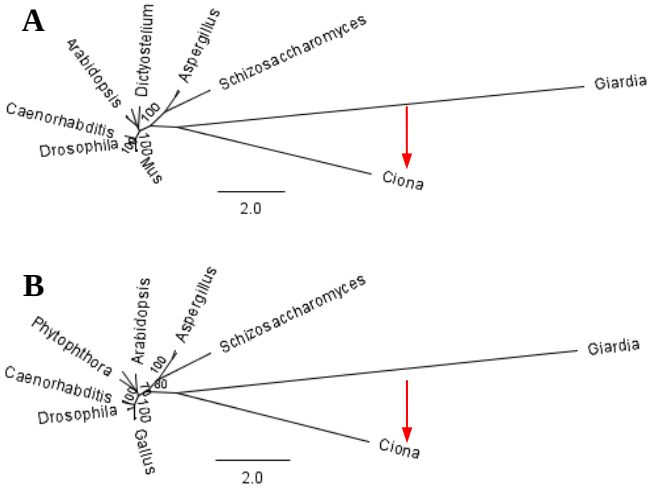

FIG. 2 UNROOTED ML TREE OF ORTHOLOGUES OF GL50803_15339 (ADAPTOR PROTEIN COMPLEX LARGE CHAIN SUBUNIT BETAA) FROM DIFFERENT SPECIES SHOWING EFFECT OF INCLUDING DIFFERENT CIONA PARALOGUES. BOOTSTRAPPING VALUES ARE SHOWN ON BRANCHES, AND THE SCALE BAR IS NUMBER OF SUBSTITUTIONS PER SITE.

A. ENSCINP00000018012 was used as Ciona orthologue. Ciona (indicated by the arrow) has been grouped with Giardia when it is expected to be grouped with other metazoans.

B. ENSCINP00000018007 was used as Ciona orthologue. Using this homologue Ciona is still approximately at the same position (Ciona branch is indicated by the arrow).

1) *Phylogenetic Analysis of Mammal Species*

Building a phylogenetic tree containing organisms from many supergroups is a difficult task and the true phylogenetic relationship between distant species is often debated. Therefore, a "simpler" phylogenetic analysis was performed on mammalian species for testing. Mammals first appeared about 225 million years ago (Kielan-Jaworowska 2007) and there is good fossil evidence for mammalian evolution with which the ESP results can be compared. This analysis here investigates if ESPs are good candidates for phylogenetic analysis over a shorter evolutionary distance compared to the entire eukaryotic tree.

Genomes of 15 mammalian species, as well as ***Anolis carolinensis*** (Anole Lizard) which serves as the outgroup were downloaded (Table 2). Each of the 50 Group A and 90 Group B ESPs was BLASTed against each species, and the homologue with the highest bit-score was recorded from each organism. The annotated transcript sequences were used for this analysis because the mammals are closely related, and the comparison of nucleotide sequences should in theory obtain better results. The 140 cDNA sequences were concatenated for each organism, and a phylogenetic tree was built using PHYML (Fig. 3). The resulting tree is almost identical to previously published and highly regarded mammalian trees (e.g. (Campbell and Lapointe ; Prasad et al. 2008; Asher, Bennett, and Lehmann 2009)), indicating that the ESPs are indeed very good for the phylogenetic analysis of species with around 200 million years of divergence.

TABLE 2 MAMMALIAN SPECIES USED IN THE PHYLOGENETIC ANALYSIS, ALL SEQUENCES ARE DOWNLOADED FROM ENSEMBL

Species	Common names
Ailuropoda melanoleuca	Panda
Bos taurus	Cow
Callithrix jacchus	Marmoset
Canis familiaris	Dog
Equus caballus	Horse
Gorilla gorilla	Gorilla
Homo sapiens	Human
Loxodonta africana	Elephant
Monodelphis domestica	Opossum
Mus musculus	Mouse
Ornithorhynchus anatinus	Platypus
Oryctolagus cuniculus	Rabbit
Pan troglodytes	Chimpanzee
Pongo pygmaeus	Orangutan
Sus scrofa	Pig

FIG. 3 PHYLOGENETIC TREE OF MAMMALIAN SPECIES. TREE WAS BUILT USING 140 CDNAS CONCATENATED FROM EACH SPECIES. HKY85 WAS THE SUBSTITUTION MATRIX. THE TREE WAS BOOTSTRAPPED 200 TIMES AND BOOTSTRAP VALUES ARE SHOWN ON THE TREE.

2) *Phylogenetic Analysis Deep Eukaryotes*

The analysis above was then performed on the 18 eukaryotic species used during ESP calculation (Han and Collins 2012). Sequences of the same species from group A and B alignments were concatenated. This eukaryotic phylogenetic tree, generated using the longest concatenated sequences to date, contained a total of 140 genes and the entire concatenated alignment was 139,625 amino acids in length. Previously Hampl *et al.* performed a similar analysis based on 143 genes and their entire concatenated alignment was 35,584 amino acids length, but this alignment suffered from a large amount of missing data (averaging 44% per taxon) (Hampl et al. 2009). Our tree was built containing more sites and less missing data and was bootstrapped 100 times (Fig. 4).

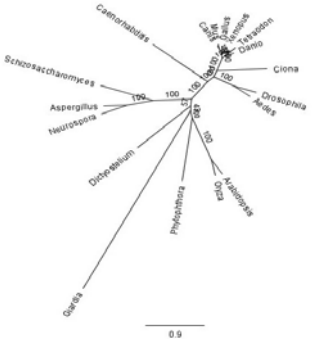

FIG. 4 UNROOTED TREE GENERATED WITH 50 GROUP A AND 50 GROUP B SEQUENCES CONCATENATED. THE TREE WAS BUILT USING WAG SUBSTITUTION MODEL WITH ESTIMATED PROPORTION OF INVARIABLE SITES (+I) AND 4 SUBSTITUTION RATE CATEGORIES AND ESTIMATED GAMMA DISTRIBUTION (+G). BOOTSTRAPPING VALUES WERE SHOWN ON BRANCHES. THE SCALE BAR IS NUMBER OF SUBSTITUTIONS PER SITE.

The animals, fungi and plants all formed monophyletic groups of with 100% bootstrap support. The bootstrap for supergroup Opisthokonta (containing animals and fungi) was moderate (57%), considering previous studies

strongly supporting that opisthokonts form a monophyletic group (Parfrey et al. 2006; Steenkamp, Wright, and Baldauf 2006). There are only four species that do not belong to supergroup Unikonta which are *Giardia*, *Phytophthora*, *Arabidopsis* and *Oryza*. The support for supergroup Unikonta (containing Opisthokonta and Amoebozoa) was low (43%). There are very few recent studies on the monophyly of unikonts (this supergroup originally proposed on that unikonts ancestrally had a single flagellum and single basal body (Cavalier-Smith 2002)). From our analysis, it is inconclusive whether Unikonta is monophyletic, due to the low bootstrap support.

Phytophthora, the only Chromalveolata species, branched closer to plants than other species from the other supergoups, with 100% bootstrap support. This is different to the tree constructed by Keeling *et al.* (Keeling 2007) which indicated Chromalveolata and Plantae were two independent supergroups. This result, however, agreed with Hampl *et al.*'s study (Hampl et al. 2009) suggesting that Chromalveolata and Archaeplastida are paraphyletic (i.e. all members along with some other unmentioned species derived from a unique common ancestor).

Giardia formed the longest branch out of all taxa, and this was also observed in the consensus networks constructed. The reason for this phenomenon we suggest is that the parasitic life style of *Giardia*, has been adapted to the constant change of the host organisms' condition, so the mutation rate of *Giardia* proteins appears faster than that of the other taxa analysed. Another tree was built with *Giardia* removed, and with the exclusion of this long branch species, the bootstrap support for every branch went up to 100% (Fig. 5).

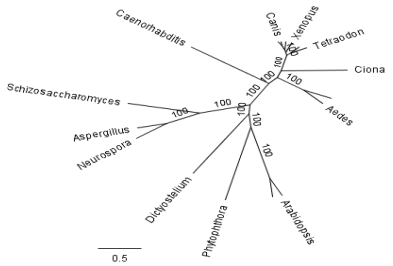

FIG. 5 UNROOTED TREE GENERATED WITH 50 GROUP A AND 50 GROUP B SEQUENCES CONCATENATED WITH GIARDIA REMOVED THE TREE WAS BUILT USING WAG SUBSTITUTION MODEL WITH ESTIMATED PROPORTION OF INVARIABLE SITES (+I) AND 4 SUBSTITUTION RATE CATEGORIES AND ESTIMATED GAMMA DISTRIBUTION (+G). BOOTSTRAPPING VALUES WERE SHOWN ON BRANCHES. THE SCALE BAR IS NUMBER OF SUBSTITUTIONS PER SITE.

Disscussion

ESPs as a group, hold interesting potential for evolutionary studies. Here a systematic way of using phylogenomics to analyse the phylogenetic relationship between eukaryotic organisms of long to medium divergence has been developed (from 1850 million to 225 million years). Our aim was to find a group of proteins suitable for the analysis, which will drastically outperform the approach of using a random set of proteins seen in most literature (e.g. (Burki et al. 2007; Hampl et al. 2009)). The assumption can be made that all ESPs have evolved from the last eukaryotic common ancestor, thus the rates of evolution over the long period until today are expected to be quite close. Multigene analysis requires two criteria: 1, Genes are present in majority of species analysed; 2, the orthology of gene in every species is clear. Using the ESP approach, the first criteria is automatically fulfilled due to the universal presence in eukaryotes. The second criteria can be challenging, as the ESP alignments were manually checked, and the ones susceptible to the problem of wrong homologue being incorporated were discarded from the analysis (i.e. the exclusion of Group C ESPs). Future research might consider discarding Group C ESPs as they can produce biased results.

The remaining Group A and B ESPs are excellent candidates for tree-building especially using the sequence concatenation method of tree-building and performed very well in resolving phylogenies among mammals. The deep eukaryotic analysis did group together all the established monophyletic groups, although there were some low bootstrap values. The conflict across the entire eukaryotic tree was expected since species were included from several supergroups. Some taxa (e.g. *Giardia* and *Phytophthora*) form a long branch on their own, these long branches could be split when more closely related organisms are included in the analysis. Generally, with more genomes becoming available, especially those of highly divergent organisms, the ESP approach was expected to produce even better results. In addition, when the long branch of *Giardia* was taken out (Fig. 5), the bootstrap value of the tree went up, again indicating that given a good model, ESPs are very good candidates for phylogenetic analyses. Resolving the root amongst the five eukaryotic supergroups is a difficult task for any research. This study definitely did not attempt to solve this problem, but to develop a method that can largely contribute towards unrevealing this enigma, when more eukaryotic genomes become available.

For future work, it would be interesting to examine if ESPs can produce good results when phylogeny of other well understood clades of eukaryotes is analysed. This will further consolidate that ESPs are valuable for phylogenetic analyses. A number of animal clades with unclear phylogenetic relationship can also be analysed using the ESP approach, e.g. the relationship between insects and other groups of arthropods (555mya of divergence) (Strausfeld and Andrew 2011). Furthermore, phylogenetics of deep branching eukaryote taxa can take place using the ESP approach for a second time, when more high coverage genomes of other early diverging and evolutionarily important eukaryotes become available (e.g. *Naegleria gruberi* has recently been sequenced (Fritz-Laylin et al. 2010) and could be a useful species to be included in the study). Inclusion of more basal eukaryotic organisms, such as excavates, chromalveolates or Rhizaria species in the analysis could answer many more questions about the relationship between the supergroups, and decipher the monophyly of unresolved groups of eukaryotes.

ACKNOWLEDGMENT

This work was funded by Health Research Council of New Zealand (HRC)-Emerging Researcher Grant 07/168.

Many thanks to Dr. Simon Hill for his help with various software.

REFERENCES

Abouheif, E., R. Zardoya, and A. Meyer. 1998. Limitations of Metazoan 18S rRNA Sequence Data: Implications for Reconstructing a Phylogeny of the Animal Kindom and Inferring the Reality of the Cambrian Explosion. Journal of Melocular Evolution 47:394-405.

Andersson, J. O., A. M. Sjogren, L. A. M. Davis, T. M. Embley, and A. J. Roger. 2003. Phylogenetic Analyses of Diplomonad Genes Reveal Frequent Lateral Gene Transfers Affecting Eukaryotes. Current Biology 13:94-104.

Asher, R. J., N. Bennett, and T. Lehmann. 2009. The New Framework for Understanding Placental Mammal Evolution. Bioessays 31:853-864.

Burki, F., K. Shalchian-Tabrizi, M. Minge, A. Skjaeveland, S. I. Nikolaev, K. S. Jakobsen, and J. Pawlowski. 2007. Phylogenomics Reshuffles the Eukaryotic Supergroups. PLos ONE 2.

Campbell, V., and F.-J. Lapointe. Retrieving a Mitogenomic Mammal Tree Using Composite Taxa. Molecular Phylogenetics and Evolution 58:149-156.

Cavalier-Smith, T. 2002. The Phagotrophic Origin of Eukaryotes and Phylogenetic Classification of Protozoa. International Journal of Systematic and Evolutionary Microbiology 52:297-354.

Drummond AJ, Ashton B, Buxton S, Cheung M, Cooper A, Duran C, Field M, Heled J, Kearse M, Markowitz S, Moir R, Stones-Havas S, Sturrock S, Thierer T, and W. A. 2011. Geneious v5.4, Available from http: // www .geneious.com/.

Fritz-Laylin, L. K., S. E. Prochnik, M. L. Ginger, J. B. Dacks, M. L. Carpenter, M. C. Field, A. Kuo, A. Paredez, J. Chapman, J. Pham, S. Shu, R. Neupane, M. Cipriano, J. Mancuso, H. Tu, A. Salamov, E. Lindquist, H. Shapiro, S. Lucas, I. V. Grigoriev, W. Z. Cande, C. Fulton, D. S. Rokhsar, and S. C. Dawson. 2010. The Genome of Naegleria Gruberi Illuminates Early Eukaryotic Versatility. Cell 140:631-642.

Guindon, S., and O. Gascuel. 2003. A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. Systematic Biology 52:696-704.

Hampl, V., L. Hug, J. W. Leigh, J. B. Dacks, B. F. Lang, A. G. B. Simpson, and A. J. Roger. 2009. Phylogenomic Analyses Support the Monophyly of Excavata and Resolve Relationships among Eukaryotic "Supergroups". Proceedings of the National Academy of Sciences of the United States of America 106:3859-3864.

Han, J., and L. Collins. 2012. Eukaryotic Signature Proteins. Journal of Proteomics and Genomic Research Paper Accepted (in Press).

Hartman, H., and A. Fedorov. 2002. The Origin of the Eukaryotic Cell: A Genomic Investigation. Proceedings of the National Academy of Sciences of the United States of America 99:1420-1425.

Hasegawa, M., H. Kishino, and T. A. Yano. 1985. Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA. Journal of Molecular Evolution 22:160-174.

Hashimoto, T., Y. Nakamura, F. Nakamura, T. Shirakura, J. Adachi, N. Goto, K. Okamoto, and M. Hasegawa. 1994. Protein Phylogeny Gives a Robust Estimation for Early Divergences of Eukaryotes - Phylogenetic Place of a Mitochondria-Lacking Protozoan, Giardia lamblia. Molecular Biology and Evolution 11:65-71.

Keeling, P. J. 2007. Deep Questions in the Tree of Life. Science 317:1875-1876.

Keeling, P. J., G. Burger, D. G. Durnford, B. F. Lang, R. W. Lee, R. E. Pearlman, A. J. Roger, and M. W. Gray. 2005. The Tree of Eukaryotes. Trends in Ecology & Evolution 20:670-676.

Kielan-Jaworowska, Z. 2007. The Beginning of the Age of Mammals. Nature 446:264-265.

Knoll, A. H., E. J. Javaux, D. Hewitt, and P. Cohen. 2006. Eukaryotic Organisms in Proterozoic Oceans. Philosophical Transactions of the Royal Society B-Biological Sciences 361:1023-1038.

Kurland, C. G., L. J. Collins, and D. Penny. 2006. Genomics and the Irreducible Nature of Eukaryote Cells. Science 312:1011-1014.

Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson, and D. G. Higgins. 2007. Clustal W and Clustal X Version 2.0. Bioinformatics 23:2947-2948.

Meyer, A., C. Todt, N. T. Mikkelsen, and B. Lieb. 2010. Fast Evolving 18S rRNA Sequences from Solenogastres (Mollusca) Resist Standard PCR Amplification and Give New Insights into Mollusk Substitution Rate Heterogeneity. Bmc Evolutionary Biology 10:Article No.: 70.

Morrison, H. G., A. G. McArthur, F. D. Gillin, S. B. Aley, R. D. Adam, G. J. Olsen, A. A. Best, W. Z. Cande, F. Chen, M. J. Cipriano, B. J. Davids, S. C. Dawson, H. G. Elmendorf, A. B. Hehl, M. E. Holder, S. M. Huse, U. U. Kim, E. Lasek-Nesselquist, G. Manning, A. Nigam, J. E. J. Nixon, D. Palm, N. E. Passamaneck, A. Prabhu, C. I. Reich, D. S. Reiner, J. Samuelson, S. G. Svard, and M. L. Sogin. 2007. Genomic Minimalism in the Early Diverging Intestinal Parasite Giardia Lamblia. Science 317:1921-1926.

Nixon, J. E. J., A. Wang, J. Field, H. G. Morrison, A. G. McArthur, M. L. Sogin, B. J. Loftus, and J. Samuelson. 2002. Evidence for Lateral Transfer of Genes Encoding Ferredoxins, Nitroreductases, NADH Oxidase, and Alcohol Dehydrogenase 3 from Anaerobic Prokaryotes to Giardia Lamblia and Entamoeba Histolytica. Eukaryotic Cell 1:181-190.

Parfrey, L. W., E. Barbero, E. Lasser, M. Dunthorn, D. Bhattacharya, D. J. Patterson, and L. A. Katz. 2006. Evaluating Support for the Current Classification of Eukaryotic Diversity. Plos Genetics 2:2062-2073.

Philippe, H. 2000. Opinion: Long Branch Attraction and Protist Phylogeny. Protist 151:307-316.

Philippe, H., and A. Adoutte. 1998. The Molecular Phylogeny of Eukaryota: Solid Facts and Uncertainties.

Prasad, A. B., M. W. Allard, E. D. Green, and N. C. S. Program. 2008. Confirming the Phylogeny of Mammals by Use of Large Comparative Sequence Data Sets. Molecular Biology and Evolution 25:1795-1808.

Simpson, A. G. B. 2003. Cytoskeletal Organization, Phylogenetic Affinities and Systematics in the Contentious Taxon Excavata (Eukaryota). International Journal of Systematic and Evolutionary Microbiology 53:1759-1777.

Simpson, A. G. B., Y. Inagaki, and A. J. Roger. 2006. Comprehensive Multigene Phylogenies of Excavate Protists Reveal the Evolutionary Positions of "Primitive" Eukaryotes. Molecular Biology and Evolution 23:615-625.

Steenkamp, E. T., J. Wright, and S. L. Baldauf. 2006. The Protistan Origins of Animals and Fungi. Molecular Biology and Evolution 23:93-106.

Strausfeld, N. J., and D. R. Andrew. 2011. A New View of Insect-Crustacean Relationships I. Inferences from Neural Cladistics and Comparative Neuroanatomy. Arthropod Structure & Development 40:276-288.